

CLUSTERING FOR MULTIVARIATE CONTINUOUS AND DISCRETE LONGITUDINAL DATA

BY ARNOŠT KOMÁREK¹ AND LENKA KOMÁRKOVÁ²

Charles University in Prague and University of Economics in Prague

Multiple outcomes, both continuous and discrete, are routinely gathered on subjects in longitudinal studies and during routine clinical follow-up in general. To motivate our work, we consider a longitudinal study on patients with primary biliary cirrhosis (PBC) with a continuous bilirubin level, a discrete platelet count and a dichotomous indication of blood vessel malformations as examples of such longitudinal outcomes. An apparent requirement is to use all the outcome values to classify the subjects into groups (e.g., groups of subjects with a similar prognosis in a clinical setting). In recent years, numerous approaches have been suggested for classification based on longitudinal (or otherwise correlated) outcomes, targeting not only traditional areas like biostatistics, but also rapidly evolving bioinformatics and many others. However, most available approaches consider only continuous outcomes as a basis for classification, or if non-continuous outcomes are considered, then not in combination with other outcomes of a different nature. Here, we propose a statistical method for clustering (classification) of subjects into a prespecified number of groups with a priori unknown characteristics on the basis of repeated measurements of several longitudinal outcomes of a different nature. This method relies on a multivariate extension of the classical generalized linear mixed model where a mixture distribution is additionally assumed for random effects. We base the inference on a Bayesian specification of the model and simulation-based Markov chain Monte Carlo methodology. To apply the method in practice, we have prepared ready-to-use software for use in R (<http://www.R-project.org>). We also discuss evaluation of uncertainty in the classification and also discuss usage of a recently proposed methodology for model comparison—the selection of a number of clusters in our case—based on the penalized posterior deviance proposed by Plummer [*Biostatistics* **9** (2008) 523–539].

Received March 2011; revised July 2012.

¹Supported by the Czech Science Foundation Grant GAČR 201/09/P077.

²Supported by the Czech Science Foundation Grant GAČR P403/12/1557.

Key words and phrases. Classification, functional data, generalized linear mixed model, multivariate longitudinal data, repeated observations.

This is an electronic reprint of the original article published by the Institute of Mathematical Statistics in *The Annals of Applied Statistics*, 2013, Vol. 7, No. 1, 177–200. This reprint differs from the original in pagination and typographic detail.

1. Introduction.

1.1. *Data and the research question.* In clinical practice multiple markers of disease progression, both continuous and discrete, are routinely gathered during the follow-up to decide on future treatment actions. Our work is motivated by data from a Mayo Clinic trial on 312 patients with primary biliary cirrhosis (PBC) conducted between 1974–1984 [Dickson et al. (1989)]. This longitudinal study had a median follow-up time of 6.3 years with a large number of clinical, biochemical, serological and histological parameters recorded for each patient. The data are available in Fleming and Harrington (1991), Appendix D, and electronically at <http://lib.stat.cmu.edu/datasets/pbcseq>.

With these data, we shall mimic a common problem from the clinical practice: at a prespecified time point from the start of follow-up we want to use the values of the markers of the disease progression to identify groups of patients with similar characteristics. That is, we want to perform a cluster analysis using the longitudinal measurements. With these motivating data, we perform a classification of patients who survived without liver transplantation the first 910 days (2.5 years) of the study ($N = 260$), the data being further referred to as PBC910. This corresponds to the practical problem outlined above, that is, clustering of patients being available at a given time point. For the purpose of this paper, the time point of 910 days was selected arbitrarily. Its choice in other application can, of course, be driven by practical or other considerations. The following markers will be considered for the cluster analysis: continuous logarithmic serum bilirubin (lbili), discrete platelet count (platelet) and dichotomous indication of blood vessel malformations (spiders); see Figure 1.

In a clinical routine usually only the last available measurements reflecting the current patient status are used to identify the prognostic groups—clusters. Clearly, such a procedure ignores the available information on the markers' evolution over time, which might be more important for reasonable classification than simply the last known state. To remedy this deficiency, we shall propose a clustering method exploiting jointly the whole history of longitudinal measurements of all considered markers which might have a different nature from being continuous to discrete, or even dichotomous.

1.2. *Basic notation and data characteristics.* Let $\mathbf{Y}_{i,r} = (Y_{i,r,1}, \dots, Y_{i,r,n_{i,r}})^\top$ denote a random vector of the longitudinal profile of the r th marker ($r = 1, \dots, R$) pertaining to the i th subject ($i = 1, \dots, N$). Further, let $\mathbf{Y}_i = (\mathbf{Y}_{i,1}^\top, \dots, \mathbf{Y}_{i,R}^\top)^\top$ be a random vector of all longitudinal measurements on the i th subject and $\mathbf{Y} = (\mathbf{Y}_1^\top, \dots, \mathbf{Y}_N^\top)^\top$ a random vector representing all available outcomes. As usual, let $y_{i,r,j}, \mathbf{y}_{i,r}, \mathbf{y}_i, \mathbf{y}$ denote the observed counterparts of corresponding upper case random variables and vectors. Throughout the paper, we will assume the independence of subjects,

TABLE 1

Data PBC910. Characteristics of the time points at which the longitudinal values of the markers for clustering were taken. For each marker r and each visit j , $n_{r,j}^*$ gives the number of available measurements, med , Q_1 and Q_3 are the median, the lower and the upper quartile of the time points in months when the measurements were taken

j	$n_{r,j}^*$	lbili ($r = 1$)		platelet ($r = 2$)			spiders ($r = 3$)		
		$t_{i,r,j}$ (months)		$t_{i,r,j}$ (months)			$t_{i,r,j}$ (months)		
		med	(Q_1-Q_3)	$n_{r,j}^*$	med	(Q_1-Q_3)	$n_{r,j}^*$	med	(Q_1-Q_3)
1	260	0.0	(0.0–0.0)	256	0.0	(0.0–0.0)	260	0.0	(0.0–0.0)
2	248	6.1	(5.9–6.7)	241	6.1	(5.9–6.7)	247	6.1	(5.9–6.8)
3	226	12.2	(11.8–12.9)	224	12.2	(11.8–12.9)	224	12.2	(11.8–12.9)
4	181	24.3	(23.8–25.3)	180	24.3	(23.8–25.2)	180	24.3	(23.8–25.2)
5	3	23.4	(23.4–26.5)	2	23.4	(23.4–23.4)	2	23.4	(23.4–23.4)

that is, independence of $\mathbf{Y}_1, \dots, \mathbf{Y}_N$. Furthermore, let $n = \sum_{i=1}^N \sum_{r=1}^R n_{i,r}$ be the total number of observations and let $t_{i,r,j}$ be the times (on a study time scale) at which the individual values $Y_{i,r,j}$ ($i = 1, \dots, N$, $r = 1, \dots, R$, $j = 1, \dots, n_{i,r}$) were taken. Finally, let $p(\cdot)$ and $p(\cdot|\cdot)$ be generic symbols for (conditional) distributions.

In the PBC910 data and our application, the number of markers R equals 3. As it is common with the longitudinal data, numbers $n_{i,r}$ of available measurements of each marker varies (between 1 and 5) across patients (median 4), leading to $n = 2734$. Further, the distribution of the time points $t_{i,r,j}$ also varies across subjects and markers; see Table 1. However, note that the fifth visit which is available for only three patients is not outlying with respect to its timing from the rest of the data set. Indeed, it only corresponds to patients with a slightly more frequent visiting schedule. In summary, our longitudinal data are heavily unbalanced and irregularly spaced in time, also containing 12 patients for whom only baseline marker values at time $t = 0$ are available. It is our aim to also classify or at least suggest a classification for those patients.

1.3. *Existing clustering methods, the need for extensions.* In the literature numerous clustering methods applicable in many different situations are available. Nevertheless, as we discuss below, none of them is applicable for our problem of clustering where each subject is represented by a set of R , in general unbalanced and irregularly sampled longitudinal profiles of markers which may have a different nature starting from continuous and ending with a dichotomous one.

1.3.1. *Classical approaches and a mixture model-based clustering.* Apart from classical approaches like hierarchical clustering or K -means method [see, e.g., Hastie, Tibshirani and Friedman (2009), Chapter 13, Johnson and

Wichern (2007), Chapter 12] and their many extensions, model-based clustering built on mixtures of parametric or even nonparametrically specified distributions assumed for random vectors $\mathbf{Y}_1, \dots, \mathbf{Y}_N$ has become quite popular in the past decade [e.g., Fraley and Raftery (2002)]. This is probably partially due to the availability of ready-to-use software like R [R Development Core Team (2012)] packages `mclust` of Fraley and Raftery (2006) for clustering based on mixtures of multivariate normal distributions, earlier versions of the `mixAK` package described by Komárek (2009) or `mixtools` of Benaglia et al. (2009), which also allows for nonparametric estimation of the mixture components. Another rapid evolution of model-based clustering algorithms also originates from their need in gene-expression data analysis [e.g., Newton and Chung (2010), Witten (2011)]. Nevertheless, classical approaches, model-based clustering based on mixtures of distributions and many other related methods are not applicable in our context.

Some of the above mentioned methods rely on distances based on a suitable metric between the observed values of underlying random vectors $\mathbf{Y}_1, \dots, \mathbf{Y}_N$ being viewed as points in the Euclidean space of a certain dimension. However, in our situation the dimension of each \mathbf{Y}_i is generally different for each subject and typically random. Hence, it is even not possible to define a common sample space needed to define a reasonable metric to calculate the distances.

For model-based methods, on the other hand, it is necessary to assume that the random vectors $\mathbf{Y}_1, \dots, \mathbf{Y}_N$ are independent and, given the classification, identically distributed according to a suitable (multivariate) distribution. Neither of these can be assumed since for typical longitudinal data (including our PBC910 data) the measurements are taken at different time points for each subject and, hence, $\mathbf{Y}_1, \dots, \mathbf{Y}_N$ are hardly identically distributed even if the number of measurements was the same for all subjects (which is also not the case for our data).

1.3.2. Clustering based on a mixture of regression models. For data where the i th subject out of N to be classified may be represented by one response random variable Y_i and a vector of possibly fixed covariates \mathbf{x}_i , several methods for clustering based on mixtures of regression models have been developed. Among the first, Quandt and Ramsey (1978) assume a two-component mixture of two normal linear regressions. Extension into a general number of components and also a practically applicable implementation is provided by Benaglia et al. (2009). A variant of the clustering based on a mixture of regressions with application to gene-expression data is given by Qin and Self (2006). A generalization, allowing also for nonnormally distributed response random variables Y_1, \dots, Y_N , is due to Grün and Leisch (2007), who consider mixtures of generalized linear models. To apply these methods for our application, the single response random variables $Y_{i,r,j}$ could play the

role of the response variables in the mixtures of regression models and the time points $t_{i,r,j}$ the role of the fixed covariates. Nevertheless, the clustering approaches based on mixtures of regression models are also ruled out in our situation since (a) we cannot assume a single parametric distribution for all response variables in the data since both continuous and discrete response variables appear in our data set, (b) each subject is in general represented by more than one pair (response, covariates).

1.3.3. *Clustering approaches for functional data and stochastic processes.*

For given $r \in \{1, \dots, R\}$, a set $\{\mathbf{Y}_{1,r}, \dots, \mathbf{Y}_{N,r}\}$ of longitudinal trajectories of the r th marker could also be viewed as a set of functional observations or, in more general, a set of realizations of a certain random process. A similar setting is also found in the applications in genomics where $\mathbf{Y}_{i,r}$ is typically a vector representing the expression curve of gene i over time. In the functional data or genomics literature, several clustering methods have been developed for situations when it is possible to assume a decomposition of each observed value into

$$(1) \quad Y_{i,r,j} = m_{i,r}(t_{i,r,j}) + \varepsilon_{i,r,j}, \quad i = 1, \dots, N, j = 1, \dots, n_i,$$

where $m_{i,r}(t)$ is either the value of the underlying random functional or the mean i th gene-expression at time t or, in general, the underlying stochastic process, and $\varepsilon_{i,r,j}$ ($i = 1, \dots, N, j = 1, \dots, n_{i,r}$) are random variables with a zero mean and either a common variance σ^2 or subject/gene specific variances σ_i^2 ($i = 1, \dots, N$). Based on this model (1), James and Sugar (2003) and Liu and Yang (2009) developed methods for the clustering of functional data. Peng and Müller (2008) proposed a distance-based clustering method and apply it to data from online auctions. For the genomics applications, Ramoni, Sebastiani and Kohane (2002) present an agglomerative clustering procedure based on the autoregressive model in equation (1). Another gene-expression clustering application based in fact on a mixture of regression models in equation (1) is provided by Ma et al. (2006). In our situation, these methods could only be applied if there is only one continuous marker available for each patient. Hence, with the PBC910 data, clustering would have to be based only on either `lbili` values or `platelet` values if it was assumed that they come from a continuous location-shift distribution. The dichotomous `spiders` values cannot be used at all.

1.3.4. *Clustering based on mixture extensions of the mixed models.* For the analysis of the continuous longitudinal data, the linear mixed model [LMM, Laird and Ware (1982)] plays a prominent role. For given $r \in \{1, \dots, R\}$, it is based on expression

$$(2) \quad \mathbf{Y}_{i,r} = \mathbf{x}_{i,r}^\top \boldsymbol{\alpha}_r + \mathbf{z}_{i,r}^\top \mathbf{b}_{i,r} + \varepsilon_{i,r}, \quad i = 1, \dots, N,$$

where $\mathbf{x}_{i,r}$ and $\mathbf{z}_{i,r}$ are vectors of fixed covariates containing the time points $t_{i,r,j}$ ($j = 1, \dots, n_{i,r}$) and possibly other factors. Further, $\boldsymbol{\alpha}_r$ is a vector of unknown regression parameters, $\mathbf{b}_{i,r}$ are i.i.d. random variables—random effects with unknown mean β_r and a covariance matrix \mathbf{D}_r , and $\boldsymbol{\varepsilon}_{i,r}$ are independent random vectors with zero mean and a covariance matrix $\boldsymbol{\Sigma}_{i,r}$. To cluster subjects based on the continuous longitudinal data, several approaches stemming from a mixture extension of the LMM (2) have been proposed in the literature. Verbeke and Lesaffre (1996) assume a normal mixture in the distribution of random effects and apply their method to clustering of growth curves, while Celeux, Martin and Lavergne (2005) consider a mixture of linear mixed models and perform clustering of gene-expression data. De la Cruz-Mesía, Quintana and Marshall (2008) proceed in a similar way, however, they replace the $\mathbf{x}_{i,r}^\top \boldsymbol{\alpha}_r + \mathbf{z}_{i,r}^\top \mathbf{b}_{i,r}$ part of (2) by a nonlinear expression in $\boldsymbol{\alpha}_r$ and $\mathbf{b}_{i,r}$.

By a suitable choice of the covariate vectors and imposing a suitable structure on the error covariance matrices $\boldsymbol{\Sigma}_i$, it is possible to use the LMM (2) also for the analysis of $R > 1$ continuous longitudinal markers and for clustering based on it as was done by Villarroel, Marshall and Barón (2009), or could be done using the model of Komárek et al. (2010) who performed the discriminant analysis, though. However, analogously to Section 1.3.3, all mentioned methods could be used for our application only if we wanted to base the clustering only on lbili and/or platelet values.

One possible strategy for clustering based on not only continuous longitudinal profiles is to use a general form of the model proposed by Booth, Casella and Hobert (2008) [equation (3) in their paper], where they assume that the (not necessarily normal) distribution of longitudinal observations of a particular marker depends on cluster-specific parameters and on a vector of random effects. Nevertheless, except for this general definition, they focus in their paper on a linear mixed model which is only applicable in situations when the observed longitudinal markers are continuous.

A specific option which allows for clustering based on a single longitudinal marker of a discrete nature is to replace the LMM (2) by a generalized linear mixed model [GLMM, e.g., Molenberghs and Verbeke (2005)] and assume a suitable mixture in the distribution of random effects [Spiessens, Verbeke and Komárek (2002)]. An example of clustering based on such a model is shown in Molenberghs and Verbeke (2005), Section 23.2. Nevertheless, it is still not possible to jointly use all three (in general, all R) markers.

1.3.5. Objectives and outline of the paper. In previous paragraphs we gave a brief overview of the most common classes of clustering approaches. We also argued that none of them are capable of exploiting jointly irregular longitudinal measurements of $R \geq 1$ markers of different nature (continuous, discrete or even dichotomous) as it is required by the PBC910 data. Even

though our overview is by no means exhaustive, we are not aware of any method that would meet such needs. For these reasons we propose a clustering method that will be built upon the multivariate extension (where the word “multivariate” points to the fact that $R \geq 1$ markers will be modeled jointly) of the GLMM with a normal mixture in the distribution of random effects proposed by Spiessens, Verbeke and Komárek (2002), [SVK], and Molenberghs and Verbeke (2005), [MV]. Not only did they model just the $R = 1$ longitudinal outcome, but they also considered only a homoscedastic normal mixture. Nevertheless, this is a rather restrictive assumption, especially in our context where each mixture component should represent one cluster. Hence, to have a better ground for clustering, the heteroscedastic mixture will be considered in our proposal. Further, in their illustrations, [SVK] and [MV] usually included at most bivariate random effects. This was probably due to the fact that as a method of estimation they exploited the maximum-likelihood through the EM algorithm which starts to be computationally troublesome for models with random effects of a higher dimension. For computational complexity implied by the multivariate extension of the mixture GLMM, but not only because of this, we shall use the Bayesian inference based on the Markov chain Monte Carlo (MCMC) simulation here.

In Section 2 we next describe the multivariate mixture generalized linear mixed model which will serve as the basis for our clustering procedure and show how to apply it to the PBC910 data. The clustering procedure will be described, and the clustering of patients from the PBC910 data will be performed in Section 3. In Section 4 we discuss the possibility of estimating a number of clusters needed in situations when this does not follow from the context. We evaluate the proposed methodology in Section 5 on a simulation study and finalize the paper by a discussion in Section 6.

2. Mixture multivariate generalized linear mixed model.

2.1. Model specification. Our proposed clustering procedure is based on a multivariate mixture generalized linear mixed model (MMGLMM). We first express the conditional mean of each response profile using a standard GLMM, that is,

$$(3) \quad h_r^{-1}\{E(Y_{i,r,j}|\boldsymbol{\alpha}_r, \mathbf{b}_{i,r})\} = \mathbf{x}_{i,r,j}^\top \boldsymbol{\alpha}_r + \mathbf{z}_{i,r,j}^\top \mathbf{b}_{i,r},$$

$$i = 1, \dots, N, r = 1, \dots, R, j = 1, \dots, n_{i,r},$$

where h_r^{-1} is the link function used to model the mean of the r th marker, $\mathbf{x}_{i,r,j}$, $\mathbf{z}_{i,r,j}$ are vectors of known covariates which may include a constant for intercept, time values in which the longitudinal observations have been taken or any other additional covariates. Further, $\boldsymbol{\alpha}_r$ is a vector of unknown regression coefficients (fixed effects) and $\mathbf{b}_{i,r}$ is a vector of random effects

for the r th response specific for the i th subject. We assume hierarchically centered GLMM [Gelfand, Sahu and Carlin (1995)] where the random effects $\mathbf{b}_{1,r}, \dots, \mathbf{b}_{N,r}$ have in general a nonzero and unknown mean, let's say β_r , $r = 1, \dots, R$; see equation (5) below. Being within the GLMM framework, we assume that for each $i = 1, \dots, N$, $r = 1, \dots, R$, $j = 1, \dots, n_{i,r}$, the conditional distribution $p(y_{i,r,j} | \phi_r, \alpha_r, \mathbf{b}_{i,r})$ belongs to an exponential family with the mean specified by (3), and possibly unknown dispersion parameter ϕ_r . In a sequel, let $\psi = (\phi^\top, \alpha^\top)^\top$, where $\alpha = (\alpha_1^\top, \dots, \alpha_R^\top)^\top$, $\phi = (\phi_1^\top, \dots, \phi_R^\top)^\top$ is the vector of GLMM related parameters.

Further, let $\mathbf{b}_i = (\mathbf{b}_{i,1}^\top, \dots, \mathbf{b}_{i,R}^\top)^\top$ be a joint vector of random effects for the i th subject ($i = 1, \dots, N$). Dependence between the R longitudinal markers of a particular subject i represented by the response vectors $\mathbf{Y}_{i,1}, \dots, \mathbf{Y}_{i,R}$ is taken into account by assuming a joint distribution for the random effect vector \mathbf{b}_i which also grounds our clustering procedure. We assume that the i th subject belongs to one of a fixed number of K clusters (see Section 4 for possible approaches to choose K if this does not follow from the context of the application at hand), each cluster with a probability $w_k = P(u_i = k | \mathbf{w})$, ($0 \leq w_k \leq 1$, $k = 1, \dots, K$, $\sum_{k=1}^K w_k = 1$), where $u_i \in \{1, \dots, K\}$ is the i th subject allocation and $\mathbf{w} = (w_1, \dots, w_K)^\top$. We further assume that the corresponding random effect vector \mathbf{b}_i follows a multivariate normal distribution with an unknown mean μ_{u_i} and a (generally nondiagonal) unknown covariance matrix \mathbf{D}_{u_i} , that is,

$$p(\mathbf{b}_i | \theta, u_i = k) = \varphi(\mathbf{b}_i; \mu_k, \mathbf{D}_k), \quad i = 1, \dots, N, k = 1, \dots, K,$$

where $\varphi(\cdot | \mu, \mathbf{D})$ is a density of the (multivariate) normal distribution with a mean μ and a covariance matrix \mathbf{D} , and $\theta = (\mathbf{w}^\top, \mu_1^\top, \dots, \mu_K^\top, \text{vec}(\mathbf{D}_1), \dots, \text{vec}(\mathbf{D}_K))^\top$ is a vector of unknown parameters related to the distribution of random effects. That is, overall, we assume a multivariate normal mixture in the distribution of random effects:

$$(4) \quad \mathbf{b}_i | \theta \stackrel{\text{i.i.d.}}{\sim} \sum_{k=1}^K w_k \mathcal{MVN}(\mu_k, \mathbf{D}_k), \quad i = 1, \dots, N.$$

With this approach, we represent the unbalanced longitudinal observations $\mathbf{Y}_1, \dots, \mathbf{Y}_N$ using a set of i.i.d. random vectors $\mathbf{b}_1, \dots, \mathbf{b}_N$ which allows us to develop a clustering procedure based on ideas of the mixture model-based clustering introduced in Section 1.3.1.

Finally, we point out that given our model, the mean effect (in a total population) of covariates included in the vectors $\mathbf{z}_{i,r,j}$, $i = 1, \dots, N$, $r = 1, \dots, R$, $j = 1, \dots, n_{i,r}$ is given by

$$(5) \quad \beta = \sum_{k=1}^K w_k \mu_k.$$

This is a vector composed of R subvectors, say, β_1, \dots, β_R , which play the role of the fixed effects for covariates included in the z -covariate vectors. Hence, for identifiability reasons, it is assumed that the vectors $\mathbf{x}_{i,r,j}$ and $\mathbf{z}_{i,r,j}$, $i = 1, \dots, N$, $r = 1, \dots, R$, $j = 1, \dots, n_{i,r}$ do not contain the same covariates.

2.2. Likelihood, Bayesian estimation. For largely computational reasons, we shall use the Bayesian inference based on the output from the MCMC simulation. To this end, the model must be specified also from a Bayesian point of view. A priori, we assume the independence between the mixture related parameters $\boldsymbol{\theta}$ and the GLMM related parameters $\boldsymbol{\psi}$. That is, the prior distribution $p(\boldsymbol{\psi}, \boldsymbol{\theta})$ factorizes as $p(\boldsymbol{\psi}, \boldsymbol{\theta}) = p(\boldsymbol{\psi}) \times p(\boldsymbol{\theta})$. The factorization of the prior is typical in generalized linear mixed models and might be justified in our application because the parameters involved in the $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ vector, respectively, express different features of the model. Specifically, for $p(\boldsymbol{\theta})$, we use a multivariate version of the classical proposal of Richardson and Green (1997) as prior distribution, and, for $p(\boldsymbol{\psi})$, we adopt classically used priors in this context [see, e.g., Fong, Rue and Wakefield (2010)]. A detailed description of the assumed form of $p(\boldsymbol{\psi}, \boldsymbol{\theta})$ is provided in Appendix A of the Supplementary Material [Komárek and Komárková (2013)] where we also explain how to choose the prior hyperparameters in order to obtain a weakly informative prior distribution.

The likelihood of the MMGLMM follows from (3) and (4):

$$(6) \quad L(\boldsymbol{\psi}, \boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\psi}, \boldsymbol{\theta}) = \prod_{i=1}^N \left(\sum_{k=1}^K w_k L_{i,k}(\boldsymbol{\psi}, \boldsymbol{\theta}) \right),$$

where

$$(7) \quad L_{i,k}(\boldsymbol{\psi}, \boldsymbol{\theta}) = \int \left\{ \prod_{r=1}^R \prod_{j=1}^{n_{i,r}} p(y_{i,r,j} | \phi_r, \boldsymbol{\alpha}_r, \mathbf{b}_{i,r}) \right\} p(\mathbf{b}_i | \boldsymbol{\theta}, u_i = k) d\mathbf{b}_i,$$

$i = 1, \dots, N, k = 1, \dots, K$

is the contribution of the i th subject to the likelihood under the assumption that the random effects are distributed according to the k th mixture component.

MCMC methods are used to generate a sample $\mathcal{S}_M = \{(\boldsymbol{\psi}^{(m)}, \boldsymbol{\theta}^{(m)}) : m = 1, \dots, M\}$ from the posterior distribution $p(\boldsymbol{\psi}, \boldsymbol{\theta} | \mathbf{y}) \propto L(\boldsymbol{\psi}, \boldsymbol{\theta}) \times p(\boldsymbol{\psi}, \boldsymbol{\theta})$. Namely, a block Gibbs algorithm is used with the Metropolis–Hastings steps for those blocks of model parameters where the normalizing constant of the full conditional distribution does not have a closed form. A well-known identifiability problem which arises from the invariance of the likelihood under permutation of the component labels is solved by applying the relabeling algorithm of Stephens (2000), which is suitable for mixture models targeted

toward clustering in particular. For details of the MCMC algorithm, refer to Appendix B of the Supplement [Komárek and Komárková (2013)].

2.3. MMGLMM for the PBC910 data. The MMGLMM for the clustering of patients included in the PBC910 data will be based on longitudinal measurements of (1) logarithmic serum bilirubin (lbili , $Y_{i,1,j}$), (2) platelet counts (platelet , $Y_{i,2,j}$) and (3) dichotomous presence of blood vessel malformations (spiders , $Y_{i,3,j}$) with assumed (1) Gaussian, (2) Poisson and (3) Bernoulli distribution, respectively. Exploration of the observed longitudinal profiles (see also Figure 1) suggests the following form of the mean struc-

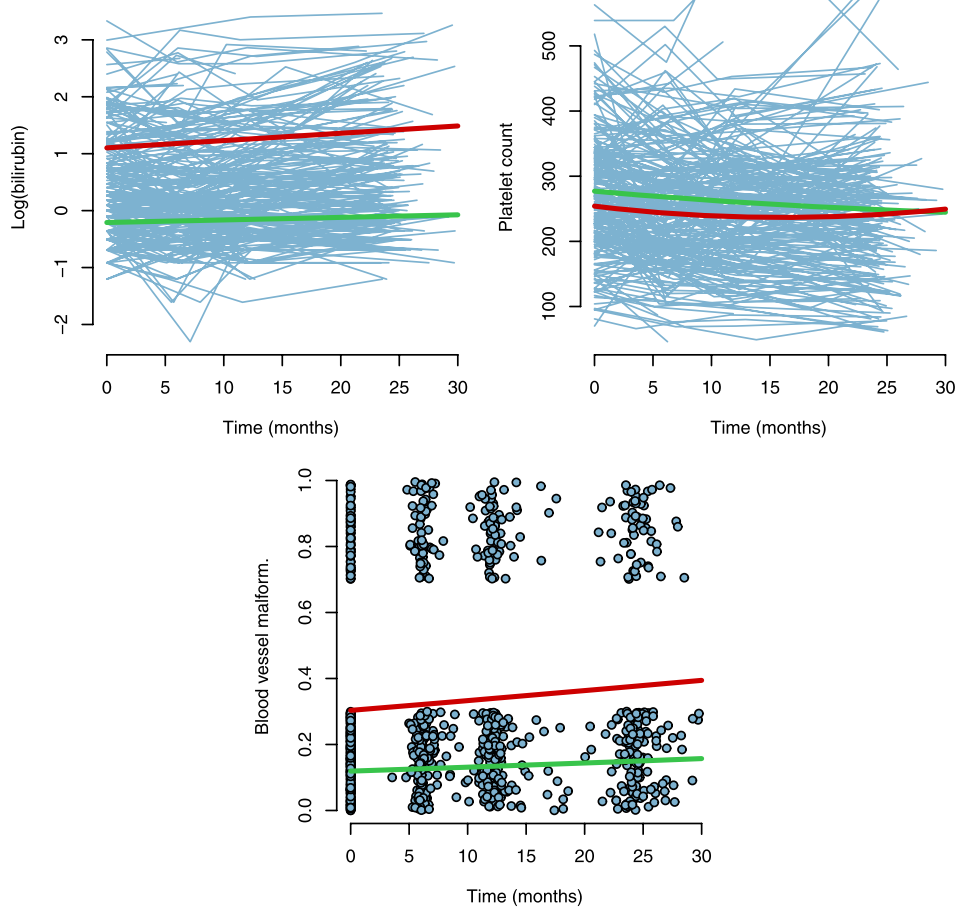


FIG. 1. *PBC910 data. Observed values of the longitudinal markers. Thick lines show cluster-specific marginal mean evolution over time based on posterior means of the mixture means μ_1 (green) and μ_2 (red). Observed values of dichotomous blood vessel malformations (spiders) are vertically jittered.*

ture (3):

$$\begin{aligned}
 \mathbb{E}(Y_{i,1,j}|\mathbf{b}_{i,1}) &= b_{i,1,1} + b_{i,1,2}t_{i,1,j}, \\
 \log\{\mathbb{E}(Y_{i,2,j}|\mathbf{b}_{i,2})\} &= b_{i,2,1} + b_{i,2,2}t_{i,2,j}, \\
 \text{logit}\{\mathbb{E}(Y_{i,3,j}|\alpha_3)\} &= b_{i,3} + \alpha_3 t_{i,3,j},
 \end{aligned}
 \tag{8}$$

$i = 1, \dots, N$, $j = 1, \dots, n_{i,r}$, $r = 1, 2, 3$, where $1 \leq n_{i,r} \leq 5$. In model (8), $t_{i,r,j}$ is the time in months from the start of the follow-up when the value of $Y_{i,r,j}$ was obtained.

In the main analysis, we will classify patients into two groups and, hence, a two component mixture ($K = 2$) will be considered in the distribution of five-dimensional random effect vector $\mathbf{b}_i = (b_{i,1,1}, b_{i,1,2}, b_{i,2,1}, b_{i,2,2}, b_{i,3})^\top$, where $b_{i,1,1}, b_{i,2,1}, b_{i,3}$ are random intercepts from the GLMM for each marker and $b_{i,1,2}, b_{i,2,2}$ are random slopes from the GLMM for the first two markers. The model also involves the fixed effect $\alpha = \alpha_3$, the slope from the logit model for the third Bernoulli response and a dispersion parameter $\phi_1 = \text{var}(Y_{i,1,j}|\mathbf{b}_{i,1})$, the residual variance from the Gaussian model for the first marker, the logarithmic bilirubin. Let $\sigma_1 = \sqrt{\phi_1}$ be the corresponding residual standard deviation. The GLMM related parameters are thus $\psi = (\sigma_1, \alpha_3)^\top$. The results that we report are based on 10,000 iterations of 1:100 thinned MCMC obtained after a burn-in period of 1000 iterations. See Appendix C of the Supplement [Komárek and Komárková (2013)] for a full Bayesian specification of the model, particular choices of the hyperparameters, an illustration of the performance of the MCMC and detailed results.

With respect to clustering, the most important parameters are the mixture weights w_1, w_2 and the mixture means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ which characterize the clusters. Their estimates taken to be the posterior means (denoted by $\hat{w}_1, \hat{w}_2, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2$, resp.) estimated from an appropriately relabeled MCMC sample are given in Table 2 together with the 95% highest posterior density credible intervals (HPD CI). The first cluster is thus characterized by a remarkably lower baseline bilirubin level and its slower increase over time compared to the second cluster. For the platelet counts, there is almost no difference between the clusters at the baseline and only a moderate difference with respect to the rate of its change, with the second cluster showing a faster decline. Finally, blood vessel malformations exhibit a higher probability in the second cluster compared to the first one. From the clinical point of view, the first cluster exhibits more favorable values and also an evolution of all three markers and, hence, it should correspond to patients with a better prognosis compared to the second cluster. We confirm this conclusion in Section 3.1 upon the classification of the individual patients.

To get a better idea of the meaning of the clusters, we used $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$ together with the posterior means of the GLMM related parameters ψ (see Table 2) and the posterior means of the mixture covariance matrices \mathbf{D}_1 and \mathbf{D}_2 (see Table 3), and calculated the estimates of the cluster specific (marginal)

TABLE 2
PBC910 data. Posterior means and 95% HPD credible intervals for mixture weights, mixture means and GLMM related parameters

	$k = 1$	$k = 2$
Parameter	$\hat{w}_1 = \mathbf{E}(w_1 \mathbf{y}) = \mathbf{0.598}$ (0.471, 0.711)	$\hat{w}_2 = \mathbf{E}(w_2 \mathbf{y}) = \mathbf{0.402}$ (0.289, 0.529)
	<i>Logarithmic bilirubin (lbili)</i>	
Intercept	-0.209	1.102
$\hat{\mu}_{k,1} = \mathbf{E}(\mu_{k,1} \mathbf{y})$	(-0.332, -0.082)	(0.828, 1.387)
Slope	0.00450	0.01281
$\hat{\mu}_{k,2} = \mathbf{E}(\mu_{k,2} \mathbf{y})$	(0.00056, 0.00818)	(0.00476, 0.02108)
Residual std. dev.	0.314	
$\hat{\sigma}_1 = \mathbf{E}(\sigma_1 \mathbf{y})$	(0.294, 0.333)	
	<i>Platelet count (platelet)</i>	
Intercept	5.58	5.46
$\hat{\mu}_{k,3} = \mathbf{E}(\mu_{k,3} \mathbf{y})$	(5.49, 5.65)	(5.35, 5.58)
Slope	-0.00567	-0.00828
$\hat{\mu}_{k,4} = \mathbf{E}(\mu_{k,4} \mathbf{y})$	(-0.00799, -0.00339)	(-0.01354, -0.00306)
	<i>Presence of blood vessel malformations (spiders)</i>	
Intercept	-4.33	-0.83
$\hat{\mu}_{k,5} = \mathbf{E}(\mu_{k,5} \mathbf{y})$	(-5.90, -2.88)	(-1.66, -0.02)
Slope	0.0280	
$\hat{\alpha}_3 = \mathbf{E}(\alpha_3 \mathbf{y})$	(0.0026, 0.0532)	

TABLE 3
PBC910 data. Standard deviations (on a diagonal) and correlations (off-diagonal elements) for each mixture component derived from the posterior means $\hat{\mathbf{D}}_1 = \mathbf{E}(\mathbf{D}_1|\mathbf{y})$ and $\hat{\mathbf{D}}_2 = \mathbf{E}(\mathbf{D}_2|\mathbf{y})$ of the mixture covariance matrices

	Intercept (lbili)	Slope (lbili)	Intercept (platelet)	Slope (platelet)	Intercept (spiders)
$k = 1$					
Intercept (lbili)	0.428	0.031	-0.282	-0.086	0.326
Slope (lbili)		0.00837	0.040	-0.214	0.100
Intercept (platelet)			0.309	-0.039	-0.042
Slope (platelet)				0.0105	0.028
Intercept (spiders)					4.02
$k = 2$					
Intercept (lbili)	0.776	-0.183	0.119	-0.139	0.171
Slope (lbili)		0.03090	-0.034	0.249	0.116
Intercept (platelet)			0.398	-0.046	-0.191
Slope (platelet)				0.0232	-0.043
Intercept (spiders)					2.42

mean longitudinal evolutions $E(Y_{.,r}, |\alpha_r, u = k) = E_{\mathbf{b}}\{E(Y_{.,r}, |\mathbf{b}, \alpha_r, u = k)\}$, $k = 1, 2$, $r = 1, 2, 3$ over time. These are plotted as green ($k = 1$) and red ($k = 2$) lines on Figure 1.

3. Clustering procedure. It follows from the decision theory for classification [see Hastie, Tibshirani and Friedman (2009), Section 2.4] that the optimal classification of the i th subject ($i = 1, \dots, N$) is to be based on the posterior component probabilities $\pi_{i,k} = P(u_i = k | \mathbf{y})$ ($k = 1, \dots, K$). In our case, they are calculated by marginalization over the posterior distribution as

$$(9) \quad \begin{aligned} \pi_{i,k} &= \int p_{i,k}(\boldsymbol{\psi}, \boldsymbol{\theta}) p(\boldsymbol{\psi}, \boldsymbol{\theta} | \mathbf{y}) d(\boldsymbol{\psi}, \boldsymbol{\theta}) = E\{p_{i,k}(\boldsymbol{\psi}, \boldsymbol{\theta}) | \mathbf{y}\} \\ &\approx \frac{1}{M} \sum_{m=1}^M p_{i,k}(\boldsymbol{\psi}^{(m)}, \boldsymbol{\theta}^{(m)}) = \hat{\pi}_{i,k}, \end{aligned}$$

where by Bayes' rule

$$(10) \quad p_{i,k}(\boldsymbol{\psi}, \boldsymbol{\theta}) = P(u_i = k | \boldsymbol{\psi}, \boldsymbol{\theta}, \mathbf{y}_i) = \frac{w_k L_{i,k}(\boldsymbol{\psi}, \boldsymbol{\theta})}{\sum_{l=1}^K w_l L_{i,l}(\boldsymbol{\psi}, \boldsymbol{\theta})},$$

$i = 1, \dots, N, k = 1, \dots, K.$

The i th subject is classically assigned to the cluster g_i for which $\hat{\pi}_{i,g_i}$ is largest among (9) [e.g., Titterton, Smith and Makov (1985), McLachlan and Basford (1988)].

In non-Bayesian applications, the clustering procedure is usually based on the values of $\hat{p}_{i,k} = p_{i,k}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\psi}}$ and $\hat{\boldsymbol{\theta}}$ are suitable estimates (e.g., maximum-likelihood) of the model parameters. Only rarely the uncertainty in the estimation of $\hat{p}_{i,k}$ expressed by evaluating their standard errors or calculating the confidence intervals is taken into account. This is probably because of the fact that $p_{i,k}(\boldsymbol{\psi}, \boldsymbol{\theta})$ depend on the model parameters in a relatively complex way. In our case of the MMGLMM, it is, for example, complicated by the necessity to integrate the GLM likelihood over the assumed distribution of the random effects [see equation (7)], which in general does not have an analytic solution.

With the Bayesian approach followed in this paper, the values of $\hat{\pi}_{i,k}$ used for clustering are the estimated posterior means of $p_{i,k}(\boldsymbol{\psi}, \boldsymbol{\theta})$ and with the MCMC-based posterior inference, we can easily evaluate also the posterior standard deviations (counterparts of the classical standard errors) or the credible intervals (counterparts of the classical confidence intervals). These can be used to incorporate an uncertainty in the classification which, for example, in the clinical setting of the PBC910 data, can serve for the identification of subjects that should undergo additional screening before their ultimate classification; see Section 3.1.

3.1. *Clustering of patients from the PBC910 data.* Classification of patients according to the maximal value of the estimated posterior mean $\hat{\pi}_{i,k}$ leads to 167 patients being classified in group 1 and 93 patients in group 2; see Figure 2. We argued in Section 2.3 that from a clinical point of view, the first group should correspond to patients with a better prognosis compared to the second group. With the PBC910 data, it is possible to confirm this conclusion since the information concerning the residual progression free survival time, defined as time till death due to liver complications or till liver transplantation, is available in the form of the classical right-censored data. We calculated Kaplan–Meier estimates of the survival probabilities based on data from patients classified in each group. These are plotted as solid lines on Figure 3. Indeed, the survival prognosis of group 1 is much better than that of group 2 with the estimated 5-year survival probability in group 1 of 0.934 compared to 0.554 in group 2, and the 10-year survival probabilities 0.679 and 0.141 in groups 1 and 2, respectively.

The fact that the posterior means $\pi_{i,k}$ of the patient specific component probabilities characterize with different certainty the probabilities of the allocation for different patients is illustrated by Figure 4. It shows the MCMC-based estimates of the posterior distributions of the first component probabilities $p_{i,1} = p_{i,1}(\boldsymbol{\psi}, \boldsymbol{\theta})$ for three selected patients who were all classified in a better prognosis group 1. For patient A, $\hat{\pi}_{i,1} = 0.990$ with a very narrow 95% HPD CI of (0.970, 1.000) and, thus, her classification in group 1 is almost certain. This is further confirmed by her progression-free survival time, which is almost 14 years. The posterior probability of belonging to group 1 is lower for patient C ($\hat{\pi}_{i,1} = 0.667$). Nevertheless, it is still twice as high as the posterior probability of belonging to group 2 ($\hat{\pi}_{i,2} = 0.333$) and, hence, it seems that patient C also belongs most likely in group 1. On the other hand, his progression-free survival time is only 3 years and 4 months (i.e., only 10 months beyond the time point at which we perform classification) and, hence, from a clinical point of view, this patient should rather be classified in group 2. In this case, the uncertainty in classification is expressed by a very wide 95% HPD CI which is (0.168, 1.000), covering the majority of the interval (0, 1).

In clinical practice, the diagnostic procedure usually proceeds in several steps, where in each step it is decided whether it is possible to classify a patient with enough certainty or whether additional examinations are needed before the ultimate classification is determined. Quite naturally, one of the diagnostic steps of such a procedure could be based on calculated credible intervals for individual component probabilities $p_{i,k} = p_{i,k}(\boldsymbol{\psi}, \boldsymbol{\theta})$. The patient would be ultimately classified in one of the considered groups only if the lower limit of the corresponding credible interval exceeds a certain threshold, say, 0.5 in the simplest case of classification into $K = 2$ groups. Applying this procedure to the PBC910 data lead to 126 patients being

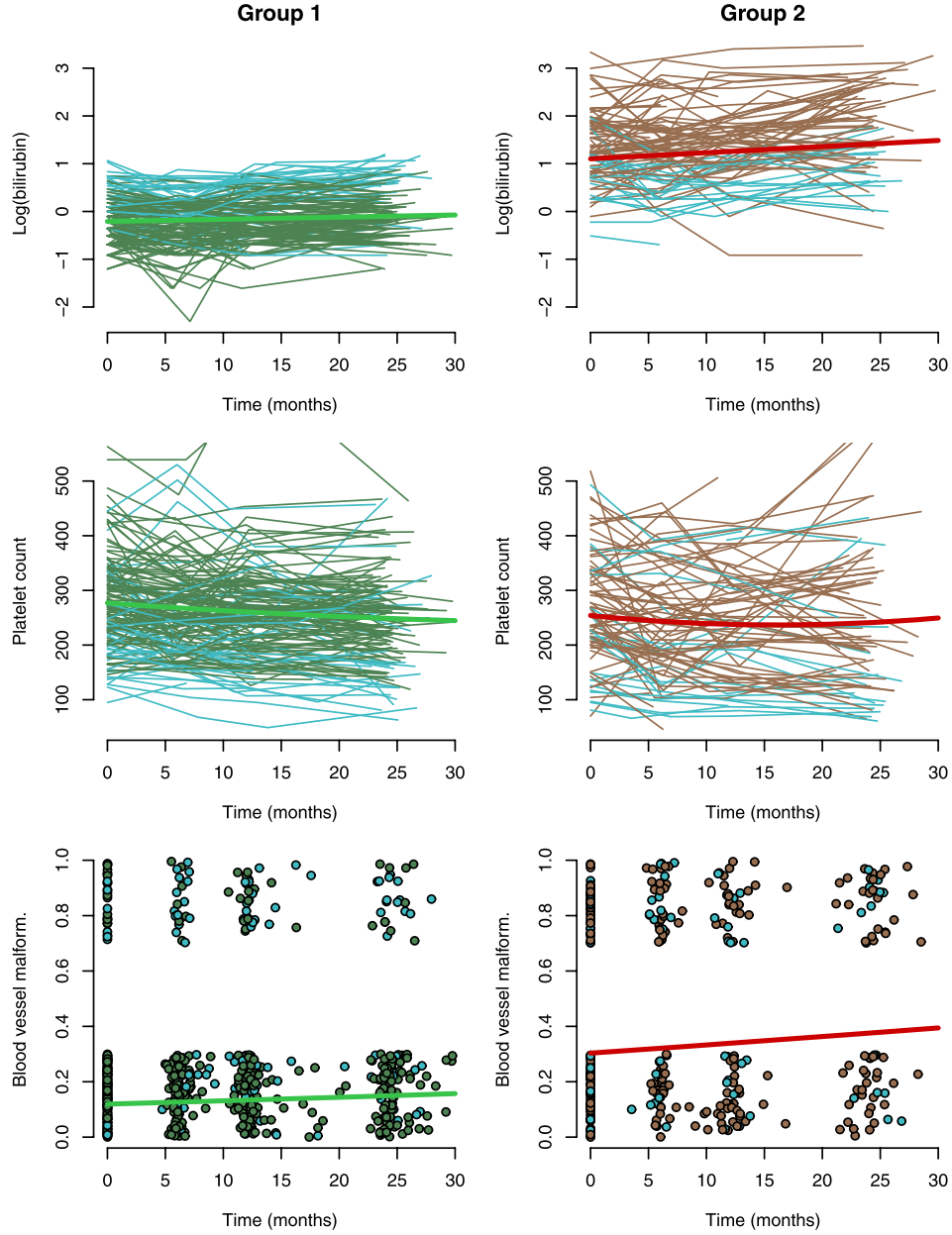


FIG. 2. *PBC910 data. Observed values of the longitudinal markers upon classification into $K = 2$ groups. The thick lines show cluster-specific marginal mean evolution over time based on posterior means of the mixture means μ_1 (green) and μ_2 (red). Observed values of dichotomous blood vessel malformations (spiders) are vertically jittered. Profiles of patients for whom the lower limit of the 95% HPD credible interval did not exceed 0.5 are drawn in light blue.*

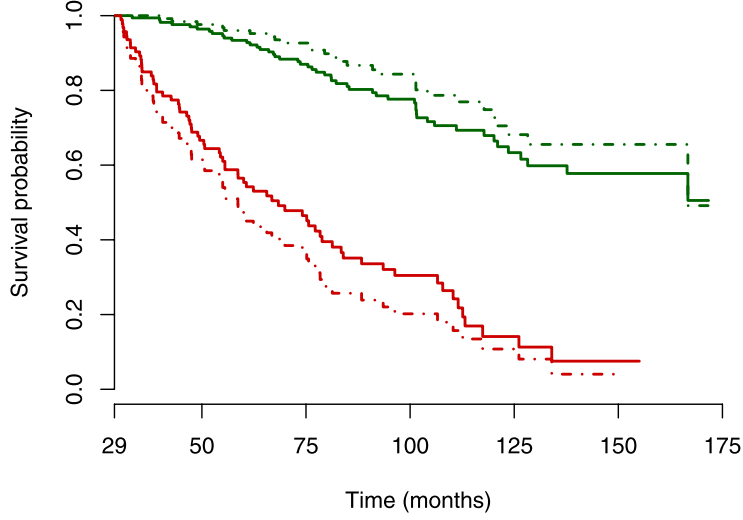


FIG. 3. *PBC910 data. Kaplan–Meier estimates of survival probability beyond 910 days in each group ($k = 1$: green, $k = 2$: red) created using the clustering procedure. Solid lines: everybody classified using the maximal value of $\hat{\pi}_{i,k}$, dotted–dashed line: only patients for whom the lower limit of the 95% HPD CI for the component probability exceeded 0.5 were classified.*

classified in group 1 and 70 in group 2. In total, 64 patients (41 and 23 originally classified in group 1 and group 2, respectively, see light blue profiles on Figure 2) remain without ultimate classification, and additional screening or examinations would have been recommended for them before making a final decision. The fact that the two groups consisting of only $126 + 70$ ultimately classified patients better reflect the progression free survival status is illustrated by the Kaplan–Meier survival curves (dotted–dashed lines on Figure 3), which are now faster, diverting with the estimated 5-year survival probability in group 1 of 0.960 compared to 0.465 in group 2, and the 10-year survival probabilities 0.748 and 0.108 in groups 1 and 2, respectively.

4. Selection of a number of clusters. Until now, we assumed that the number of clusters, K , was known in advance and fixed. In medical applications where the found clusters are expected to correspond to certain prognostic groups, this is quite a reasonable assumption. Nevertheless, in many other situations, the number of clusters should rather be inferred from the data themselves.

With our approach, the selection of a number of clusters corresponds to the selection of a number of mixture components in the underlying distribution of the random effects of the MMGLMM. This can also be viewed

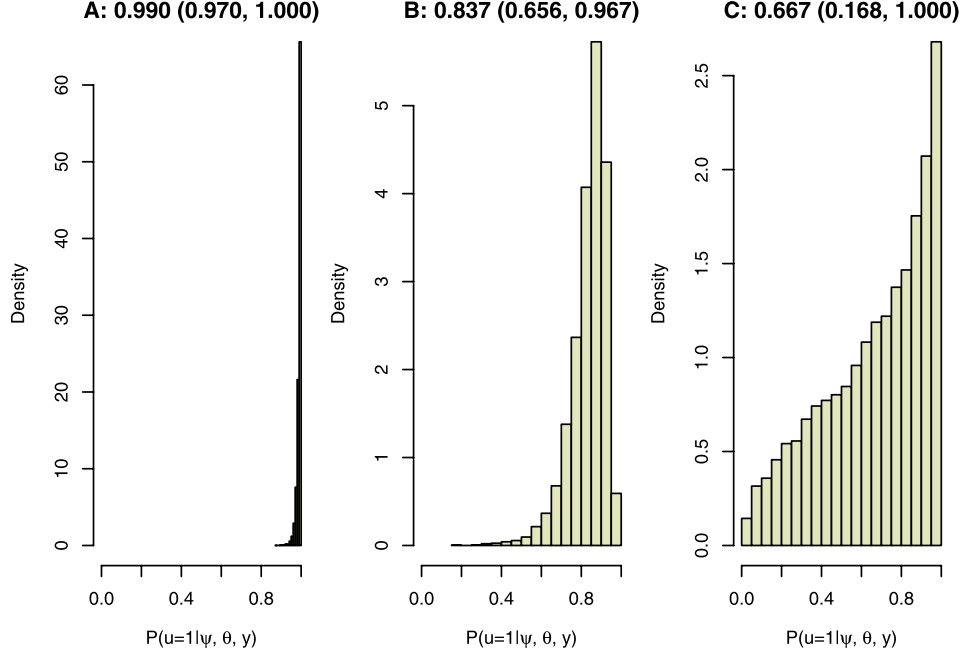


FIG. 4. *PBC910 data. Histograms of sampled values of component probabilities $p_{i,1}(\psi, \theta)$ for three selected patients. Above the plot: estimated posterior mean $\hat{\pi}_{i,1}$ and the 95% HPD CI for $p_{i,1} = p_{i,1}(\psi, \theta)$.*

as a problem of model selection or model comparison. Nevertheless, as described, for example, in McLachlan and Peel (2000), Chapter 6, the model comparison in the mixture setting is complicated by the fact that the classical regularity conditions do not hold. For this reason, use of various sorts of information criteria is predominantly preferred to classical testing in most frequentist applications of the mixture models. In particular, the Bayesian information criterion [BIC, Schwarz (1978)] proved to be a useful tool for selecting the number of mixture components [Dasgupta and Raftery (1998), Fraley and Raftery (2002), Hennig (2004), De la Cruz-Mesía, Quintana and Marshall (2008) and many others].

In Bayesian statistics, the Bayes factors [Kass and Raftery (1995)], for which the BIC is an approximation, are widely recognized as a tool of model selection. However, as pointed out by Plummer (2008), the Bayes factors have some practical limitations. First, they cannot be routinely calculated from the MCMC output. Second, they are numerically unstable when proper, but weakly informative diffused priors (as it is in our case) are used. In Bayesian applications, the deviance information criterion [DIC, Spiegelhalter et al. (2002)] seems to be the most widely used concept of model selection in the past decade. Nevertheless, DIC in the mixture context lacks

the theoretical foundations and its use remains controversial [see comments and rejoinder on Celeux et al. (2006)]. For these reasons, Plummer (2008) suggested basing criteria for model choice on penalized loss functions and cross-validating arguments. For mixture models, in particular, he suggested using the penalized expected deviance (PED). Since then, it has been successfully exploited in several applications [Komárek (2009), Cabral, Lachos and Madruga (2012), De la Fé Rodríguez et al. (2011)], and we shall use it as well to choose the number of clusters.

The penalized expected deviance is defined as

$$(11) \quad \text{PED} = \mathbb{E}\{D(\boldsymbol{\psi}, \boldsymbol{\theta})|\mathbf{y}\} + p_{\text{opt}},$$

where $D(\boldsymbol{\psi}, \boldsymbol{\theta}) = -2\log\{L(\boldsymbol{\psi}, \boldsymbol{\theta})\}$ is the observed data deviance of the model, and its posterior mean $\mathbb{E}\{D(\boldsymbol{\psi}, \boldsymbol{\theta})|\mathbf{y}\}$ (expected deviance) is easily estimated from the MCMC sample. Further, the p_{opt} part of equation (11) is the penalty term called optimism, which can be estimated by using the two parallel MCMC chains and importance sampling [Plummer (2008)].

4.1. *Number of clusters in the PBC910 data.* Table 4 shows calculated values of the penalized expected deviance for the PBC910 data and models with $K = 1, 2, 3, 4$ clusters. It shows that the two-component model fits the data clearly better than a model with just a single cluster. On the other hand, the three clusters are already too much for these data. Even though the expected deviance of the three-component model is lower than that of the two-component model, the decrease of the expected deviance is overcome by the penalization for the additional component.

The conclusion that the third cluster is redundant for our application is also supported by the fact that when a three-component model is fitted to the PBC910 data, the two components almost coincide with the mixture components from the $K = 2$ model and the estimated weight of the additional component is only $\hat{w}_3 = 0.021$, and only three patients are allocated here using the rule based on a maximal value of the posterior component probability.

TABLE 4
*PBC910 data. Penalized expected deviance for
models with $K = 1, 2, 3, 4$ clusters*

K	PED	$\hat{\mathbb{E}}\{D(\boldsymbol{\psi}, \boldsymbol{\theta}) \mathbf{y}\}$	\hat{p}_{opt}
1	14,277.9	14,241.8	36.1
2	14,164.1	14,088.3	75.8
3	14,183.1	14,057.1	126.0
4	22,405.1	17,244.4	5160.8

5. A simulation study.

5.1. *Simulation setup.* The setup of the simulation study was motivated by the PBC910 application and the data were generated according to the model (8). For each subject i and each marker r , $n_{i,r} = 4$ visit times were generated, with the first visit time being equal to 0 and the remaining three visit times being generated from uniform distributions on intervals (170, 200), (350, 390), (710, 770) days, respectively. The covariate $t_{i,r,j}$ in (8) was the visit time in months. The GLMM related parameters were equal: $\sigma_1 = \sqrt{\phi_1} = 0.3$, $\alpha_3 = 0.05$, and we tried two values for the true number of clusters: $K = 2$ and $K = 3$. In the two-cluster data, both mixture weights were rather high: $\mathbf{w} = (0.6, 0.4)^\top$, whereas in the three-cluster data, a small third component with $w_3 = 0.06$ was created by splitting the second component of the two-cluster setting, leading to the weights $\mathbf{w} = (0.60, 0.34, 0.06)^\top$. To make the differences between the clusters less obvious, not all elements of the mixture means varied across the clusters; see Table 5. Namely, in data with $K = 2$, both clusters shared the same value of the mean slope of the Gaussian response and also the same value of the mean intercept of the Poisson response. There were also some differences introduced across the mixture covariance matrices; see Table D.1 in the Supplement [Komárek and Komárková (2013)]. Example data sets generated according to considered simulation settings are also shown on Figures D.1 and D.2 of the Supplement.

To examine the performance of our method in situations when there is misspecification in the random effects distribution, we simulated data not only under the normal distribution of random effects, but also under the shifted-scaled multivariate t -distribution with five degrees of freedom (MVT₅). For each setting (K , distribution of random effects), we tried three values of sample sizes with total numbers of subjects being 50, 100, 200. For each setting and each sample size, 100 data sets were generated. The posterior inference for each data set was based on 10,000 iterations of 1:100 thinned MCMC obtained after a burn-in period of 1000 iterations.

5.2. *Parameter estimates.* The left block of Table 5 shows the square roots of the mean squared errors (MSE) in the estimation of the most important model parameters (mixture weights and means, GLMM fixed effects) provided that the number of clusters, K , is correctly specified. As parameter estimates, we considered the posterior means, and for a particular parameter the reported MSE is the average MSE over the parameter values from all K components. Detailed results also providing the bias and the standard deviations of the posterior means are given in Tables D.3—D.6 of the Supplement [Komárek and Komárková (2013)].

With normally distributed random effects, the posterior means of model parameters seem to provide consistent estimates of the model parameters. The same can also be concluded when the true distribution of random effects

TABLE 5

Simulation study: (a) proportions of models selected with $K = 1, 2, 3, 4$ using the PED criterion; (b) total classification error rate from a model with correctly specified K ; (c) true values of mixture weights and mixture means; (d) square roots of the mean squared errors (MSE), where the calculated MSE is based on posterior means as parameter estimates. For each parameter, the reported MSE is the average MSE over the K mixture components. The N gives the true number of subjects in each cluster

Setting	w	$\mu_{*,1}$	$\mu_{*,2}$	$\mu_{*,3}$	$\mu_{*,4}$	$\mu_{*,5}$	α_3	Classif. error rate (%)	Proportion (%) of models selected with				
									1	2	K	3	4
$K = 2$				<i>True values</i>									
	0.600	0.000	0.0100	5.00	−0.0050	−3.00	0.050						
	0.400	1.000	0.0100	5.00	−0.0200	−1.00							
				<i>Square root of the MSE</i>									
Normal													
$N = (30, 20)$	0.169	0.237	0.0048	0.12	0.0039	1.84	0.029	15.8	67	33	0	0	
(60, 40)	0.064	0.150	0.0034	0.09	0.0018	0.86	0.018	7.8	15	85	0	0	
(120, 80)	0.037	0.066	0.0022	0.05	0.0010	0.46	0.014	5.8	1	99	0	0	
MVT ₅													
$N = (30, 20)$	0.209	0.564	0.0134	0.36	0.0051	3.33	0.029	20.6	64	35	1	0	
(60, 40)	0.128	0.433	0.0103	0.47	0.0035	2.17	0.020	10.8	17	76	7	0	
(120, 80)	0.102	0.378	0.0049	0.16	0.0020	1.47	0.013	8.7	0	82	18	0	
$K = 3$				<i>True values</i>									
	0.600	0.000	0.0100	5.00	−0.0050	−3.00	0.050						
	0.340	1.000	0.0100	5.00	−0.0200	−1.00							
	0.060	1.300	−0.0300	5.50	0.0000	−2.00							
				<i>Square root of the MSE</i>									
Normal													
$N = (30, 17, 3)$	0.154	0.444	0.0204	0.33	0.0073	5.04	0.027	26.5	80	18	2	0	
(60, 34, 6)	0.088	0.360	0.0165	0.26	0.0054	3.07	0.018	17.4	36	50	14	0	
(120, 68, 12)	0.048	0.260	0.0126	0.19	0.0034	1.58	0.015	10.1	2	59	38	1	
MVT ₅													
$N = (30, 17, 3)$	0.113	0.563	0.0237	0.52	0.0078	3.99	0.021	23.8	66	31	3	0	
(60, 34, 6)	0.122	0.424	0.0211	0.31	0.0058	2.66	0.021	18.8	24	59	17	0	
(120, 68, 12)	0.056	0.353	0.0196	0.22	0.0048	1.56	0.013	9.7	2	44	50	4	

is MVT_5 . Nevertheless, not surprisingly, the convergence of the parameter estimates to their true values is in general slower in this case.

5.3. Classification error rates. With respect to classification, one of the most important measures is the classification error rate reported in the ninth column of Table 5. More detailed results, also showing conditional (given the cluster) classification error rates, are shown in Table D.2 of the Supplement [Komárek and Komárková (2013)]. Among other things, we point out that even with incorrectly specified distribution of random effects, the classification error rates do not differ considerably from the error rates obtained with data for which the random effects distribution was correctly specified, and even with a moderate sample size of 200 subjects, the achieved error rate is as low as 10%.

5.4. Selection of a number of clusters. Performance of the penalized expected deviance as a tool for the selection of a number of clusters is illustrated by the right block of Table 5. For each simulated data set, we estimated four models, each of them under the assumption of a different number of clusters, namely, $K = 1, 2, 3, 4$, and calculated the corresponding PED values. Table 5 shows proportions of data sets for which a particular number of clusters was selected by minimizing the PED. For each simulation setting, bold numbers indicate proportions of data sets with a correctly selected number of clusters.

For data sets composed of two clusters both having rather high weights, the probability of a selection of a correct model increases with the sample size, practically reaching 100% with $N = 200$, and normally distributed random effects. The situation is slightly worse when the true distribution of random effects is MVT_5 . Nevertheless, the results are also rather satisfactory in this case.

When there were three clusters present, with one of the clusters having a rather small weight of 0.06, the probability of a correct selection of a number of clusters also increases with the sample sizes (for both normally and MVT_5 distributed random effects). However, it is much slower than in the case of two clusters of almost equal size. Nevertheless, we point out that already for the moderate sample size of 200 subjects, in 97% and 94% of cases, respectively, the number of clusters indicated by the value of PED differs by at most one from the correct value of three.

6. Discussion. In this paper we have proposed a method for classification of subjects on the basis of longitudinal measurements of several outcomes of a different nature which, according to the best of our knowledge, has not been considered in the literature yet. The clustering procedure relies on a classical GLMM specified for each marker, whereas possible dependence

across the values of different markers is captured by specifying a joint distribution of all random effects. In contrast to a classical assumption, we assume a normal mixture in the random effects distribution which is the core classification component of the proposed model. Although other choices of the basis distributions could be considered as well, we would like to stress the fact that with the Bayesian approach used here, the normality assumption only corresponds to a particular choice of just one component of the overall prior distribution which is updated by the data. For example, the posterior predictive distribution for random effects is given by

$$p_{\text{pred}}(\mathbf{b}) = \int \left\{ \sum_{k=1}^K w_k \varphi(\mathbf{b}; \boldsymbol{\mu}_k, \mathbf{D}_k) \right\} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta},$$

which in general is no longer a Gaussian mixture. In a similar way, the posterior distribution enters the calculation of the posterior component probabilities (9) while taking into account the uncertainty in the specification of the distribution of random effects. Last but not the least, the simulation study also suggests that, at least in situations when the true random effects distribution is symmetric with heavier tails, assuming a priori a normal random effects distribution does not have any crude impact on classification error rates.

Being within the Bayesian framework, it is also relatively easy to calculate not only point estimates of the individual component probabilities, but also corresponding credible intervals which can be subsequently used to evaluate uncertainty in the pertinence of a particular subject in a specific cluster. Such uncertainty is only rarely evaluated in similar situations. Finally, we adapted recently published methodology for model comparison based on the concept of a penalized expected deviance to explore the optimal number of clusters.

Further, we point out that even though we classified in this paper only those subjects who were also used to draw the posterior inference on the model parameters, our procedure can also be used to classify a new subject with a value of observed longitudinal markers equal to \mathbf{y}_{new} and unknown allocation u_{new} . Indeed, given the posterior sample \mathcal{S}_M from $p(\boldsymbol{\psi}, \boldsymbol{\theta}|\mathbf{y})$, the component probabilities $p_{\text{new},k}(\boldsymbol{\psi}^{(m)}, \boldsymbol{\theta}^{(m)})$ and estimated posterior component probabilities $\hat{\pi}_{\text{new},k}$ ($k = 1, \dots, K$) can be calculated using expressions (10) and (9), and then used for classification. Finally, it is worth mentioning that discriminant analysis, where a training data set with known cluster allocation is available, would also be possible with only slightly modified methodology where separate posterior samples would be drawn using the data from the various clusters and then used to calculate the component probabilities.

For practical analysis, we extended the R package `mixAK` [Komárek (2009)] to cover the methodology proposed in this paper. The package is freely available from the *Comprehensive R Archive Network* at <http://cran.r-project.org/>.

Acknowledgments. We thank Mr. Steven Del Riley for his help with English corrections of the manuscript. Last but not least, we thank two anonymous referees, an Associate Editor and the Editor for very detailed and stimulative comments to the earlier versions of this manuscript that led to its considerable improvement.

SUPPLEMENTARY MATERIAL

Appendices (DOI: [10.1214/12-AOAS580SUPP](https://doi.org/10.1214/12-AOAS580SUPP); .pdf). The pdf file contains (A) more detailed description of the assumed prior distribution for model parameters, giving also some guidelines for the selection of the hyperparameters to achieve a weakly informative prior distribution; (B) more details on the posterior distribution and the sampling MCMC algorithm; (C) additional information to the analysis of the Mayo Clinic PBC data; (D) more detailed results of the simulation study.

REFERENCES

- BENAGLIA, T., CHAUVEAU, D., HUNTER, D. R. and YOUNG, D. (2009). Mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software* **32** 1–29.
- BOOTH, J. G., CASELLA, G. and HOBERT, J. P. (2008). Clustering using objective functions and stochastic search. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 119–139. [MR2412634](#)
- CABRAL, C. R. B., LACHOS, V. H. and MADRUGA, M. R. (2012). Bayesian analysis of skew-normal linear mixed models with heterogeneity in the random-effects population. *J. Statist. Plann. Inference* **142** 181–200. [MR2827140](#)
- CELEUX, G., MARTIN, O. and LAVERGNE, C. (2005). Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Stat. Model.* **5** 243–267. [MR2210736](#)
- CELEUX, G., FORBES, F., ROBERT, C. P. and TITTERINGTON, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Anal.* **1** 651–673 (electronic). [MR2282197](#)
- DASGUPTA, A. and RAFTERY, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *J. Amer. Statist. Assoc.* **93** 294–302.
- DE LA CRUZ-MESÍA, R., QUINTANA, F. A. and MARSHALL, G. (2008). Model-based clustering for longitudinal data. *Comput. Statist. Data Anal.* **52** 1441–1457. [MR2422747](#)
- DE LA FÉ RODRÍGUEZ, P. Y., CODDENS, A., DEL FAVA, E., ABRAHANTES, J. C., SHKEDY, Z., MARTIN, L. O. M., MUÑOZ, E. C., DUCHATEAU, L., COX, E. and GODDEERIS, B. M. (2011). High prevalence of F4+ and F18+Escherichia coli in Cuban piggeries as determined by serological survey. *Tropical Animal Health and Production* **43** 937–946.
- DICKSON, E. R., GRAMBSCH, P. M., FLEMING, T. R., FISHER, L. D. and LANGWORTHY, A. (1989). Prognosis in primary biliary cirrhosis: Model for decision making. *Hepatology* **10** 1–7.
- FLEMING, T. R. and HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York. [MR1100924](#)
- FONG, Y., RUE, H. and WAKEFIELD, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics* **11** 397–412.

- FRALEY, C. and RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631. [MR1951635](#)
- FRALEY, C. and RAFTERY, A. E. (2006). MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Technical Report 504, Dept. Statistics, Univ. Washington.
- GELFAND, A. E., SAHU, S. K. and CARLIN, B. P. (1995). Efficient parameterisations for normal linear mixed models. *Biometrika* **82** 479–488. [MR1366275](#)
- GRÜN, B. and LEISCH, F. (2007). Fitting finite mixtures of generalized linear regressions in R. *Comput. Statist. Data Anal.* **51** 5247–5252. [MR2370869](#)
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York. [MR2722294](#)
- HENNIG, C. (2004). Breakdown points for maximum likelihood estimators of location-scale mixtures. *Ann. Statist.* **32** 1313–1340. [MR2089126](#)
- JAMES, G. M. and SUGAR, C. A. (2003). Clustering for sparsely sampled functional data. *J. Amer. Statist. Assoc.* **98** 397–408. [MR1995716](#)
- JOHNSON, R. A. and WICHERN, D. W. (2007). *Applied Multivariate Statistical Analysis*, 6th ed. Pearson Prentice Hall, Upper Saddle River, NJ. [MR2372475](#)
- KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.
- KOMÁREK, A. (2009). A new R package for Bayesian estimation of multivariate normal mixtures allowing for selection of the number of components and interval-censored data. *Comput. Statist. Data Anal.* **53** 3932–3947. [MR2744295](#)
- KOMÁREK, A. and KOMÁRKOVÁ, L. (2013). Supplement to “Clustering for multivariate continuous and discrete longitudinal data.” DOI:[10.1214/12-AOAS580SUPP](#).
- KOMÁREK, A., HANSEN, B. E., KUIPER, E. M. M., VAN BUUREN, H. R. and LESAFFRE, E. (2010). Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Stat. Med.* **29** 3267–3283. [MR2758719](#)
- LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38** 963–974.
- LIU, X. and YANG, M. C. K. (2009). Simultaneous curve registration and clustering for functional data. *Comput. Statist. Data Anal.* **53** 1361–1376. [MR2657097](#)
- MA, P., CASTILLO-DAVIS, C. I., ZHONG, W. and LIU, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Res.* **34** 1261–1269.
- McLACHLAN, G. J. and BASFORD, K. E. (1988). *Mixture Models: Inference and Applications to Clustering. Statistics: Textbooks and Monographs* **84**. Dekker, New York. [MR0926484](#)
- McLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. Wiley, New York. [MR1789474](#)
- MOLENBERGHS, G. and VERBEKE, G. (2005). *Models for Discrete Longitudinal Data*. Springer, New York. [MR2171048](#)
- NEWTON, M. A. and CHUNG, L. M. (2010). Gamma-based clustering via ordered means with application to gene-expression analysis. *Ann. Statist.* **38** 3217–3244. [MR2766851](#)
- PENG, J. and MÜLLER, H.-G. (2008). Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *Ann. Appl. Stat.* **2** 1056–1077. [MR2516804](#)
- PLUMMER, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics* **9** 523–539.
- QIN, L.-X. and SELF, S. G. (2006). The clustering of regression models method with applications in gene expression data. *Biometrics* **62** 526–533. [MR2236835](#)

- QUANDT, R. E. and RAMSEY, J. B. (1978). Estimating mixtures of normal distributions and switching regressions. *J. Amer. Statist. Assoc.* **73** 730–738. [MR0521324](#)
- R Development Core Team. (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at <http://www.R-project.org>.
- RAMONI, M. F., SEBASTIANI, P. and KOHANE, I. S. (2002). Cluster analysis of gene expression dynamics. *Proc. Natl. Acad. Sci. USA* **99** 9121–9126 (electronic). [MR1909705](#)
- RICHARDSON, S. and GREEN, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B* **59** 731–792. [MR1483213](#)
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. [MR0468014](#)
- SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 583–639. [MR1979380](#)
- SPIESSENS, B., VERBEKE, G. and KOMÁREK, A. (2002). A SAS-macro for the classification of longitudinal profiles using mixtures of normal distributions in nonlinear and generalised linear mixed models. Technical Report, Biostatistical Center, Catholic Univ. Leuven, Leuven.
- STEPHENS, M. (2000). Dealing with label switching in mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62** 795–809. [MR1796293](#)
- TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, Chichester. [MR0838090](#)
- VERBEKE, G. and LESAFFRE, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *J. Amer. Statist. Assoc.* **91** 217–221.
- VILLARROEL, L., MARSHALL, G. and BARÓN, A. E. (2009). Cluster analysis using multivariate mixed effects models. *Stat. Med.* **28** 2552–2565. [MR2750308](#)
- WITTEN, D. M. (2011). Classification and clustering of sequencing data using a Poisson model. *Ann. Appl. Stat.* **5** 2493–2518. [MR2907124](#)

FACULTY OF MATHEMATICS AND PHYSICS
CHARLES UNIVERSITY IN PRAGUE
SOKOLOVSKÁ 83
CZ-186 75, PRAHA 8
CZECH REPUBLIC
E-MAIL: arnost.komarek@mff.cuni.cz
URL: <http://msekc.karlin.mff.cuni.cz/~komarek>

FACULTY OF MANAGEMENT
UNIVERSITY OF ECONOMICS IN PRAGUE
JAROŠOVSKÁ 1117
CZ-377 01, JINDŘICHŮV HRADEC
CZECH REPUBLIC
E-MAIL: komarkol@fm.vse.cz